REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
26-04-2011	Article	APR 2011 - MAY 2011

4. TITLE AND SUBTITLE

Towards Interpretive Models for 2-D Processing of Speech

5a. CONTRACT NUMBER

FA8720-05-C-0002

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Tianyu T. Wang and Thomas F. Quatieri

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force
Rome Research Site
525 Brooks Road
Rome, NY 13441-4114

10. SPONSOR/MONITOR'S ACRONYM(S)

AFRL/RIEC

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Two-dimensional (2-D) processing of speech has recently been explored as an alternative representational framework that explicitly analyzes temporal, spectral, and joint spectrotemporal energy fluctuations or "modulations" present in time-frequency distributions (e.g., in the spectrogram or auditory spectrogram). This paper considers 2-D Fourier analysis of local time-frequency regions of wideband spectrograms, a representation referred to as the (wideband) Grating Compression Transform (WGCT). We develop frequency dependent models of speech signals in the WGCT context related to speech production characteristics, building on previous work in modeling narrowband-based GCT representations. Model evaluation through simulations and error analysis is performed. Comparison shows the model effectiveness, and important distinctions, including "dual" behavior, between the wide and narrowband models. Our results motivate a novel taxonomy of speech signal behavior for use as an interpretative framework (i.e., in relation to speech production characteristics) for 2-D processing of speech using the GCT and potentially other 2-D approaches and time-frequency distributions. We demonstrate the ability of the model to represent real speech content through using demodulation techniques for analysis/synthesis of wideband spectrograms and co-channel speaker separation using prior pitch information.

15. SUBJECT TERMS

2-D processing of speech, Grating Compression Transform, wideband spectrogram, spectrogram reconstruction, co-channel speaker separation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE	SAR	15	Zach Sweet
U	U	U			19b. TELEPHONE NUMBER *(include area code)* 781-981-5997

JA-18505

Towards Interpretive Models for 2-D Processing of Speech

Tianyu T. Wang, *Student Member, IEEE*, and Thomas F. Quatieri, *Fellow, IEEE*

***Abstract*— Two-dimensional (2-D) processing of speech has recently been explored as an alternative representational framework that explicitly analyzes temporal, spectral, and joint spectrotemporal energy fluctuations or "modulations" present in time-frequency distributions (e.g., in the spectrogram or auditory spectrogram). This paper considers 2-D Fourier analysis of local time-frequency regions of wideband spectrograms, a representation referred to as the (wideband) Grating Compression Transform (WGCT). We develop frequency-dependent models of speech signals in the WGCT context related to speech production characteristics, building on previous work in modeling narrowband-based GCT representations. Model evaluation through simulations and error analysis is performed. Comparison shows the model effectiveness, and important distinctions, including "dual" behavior, between the wide and narrowband models. Our results motivate a novel *taxonomy* of speech signal behavior for use as an interpretative framework (i.e., in relation to speech production characteristics) for 2-D processing of speech using the GCT and potentially other 2-D approaches and time-frequency distributions. We demonstrate the ability of the model to represent real speech content through using demodulation techniques for analysis/synthesis of wideband spectrograms and co-channel speaker separation using prior pitch information.**

***Index Terms*—2-D processing of speech, Grating Compression Transform, wideband spectrogram, spectrogram reconstruction, co-channel speaker separation**

I. Introduction

Two-dimensional (2-D) processing of speech has recently been explored as an alternative representational approach that explicitly analyzes temporal, spectral, and joint spectrotemporal energy fluctuations or "modulations" present in time-frequency distributions (e.g., in the spectrogram/auditory spectrogram). Examples of this include auditory models [1][2], the modulation spectrogram [3], and our previous work in 2-D Fourier analysis of spectrograms [4][5][6][7]. Though these representations have been interpreted implicitly using data-driven techniques [2] or analytically in relation to modulation theory [8], they have nonetheless been difficult to interpret from a parametric perspective in relation to speech-specific characteristics (e.g., pitch) [9]. The aim of this work and our previous work in [10] is to provide a speech-based interpretive framework for the concept of "modulation" in 2-D processing of speech.

In [10], we developed speech-specific signal models for localized 2-D Fourier analysis of the *narrowband* spectrogram [7], a representation referred to as the (narrowband) Grating Compression Transform (NGCT). More generally, it is of interest to apply 2-D Fourier analysis to time-frequency distributions that can be viewed as mixtures of both narrowband and *wideband* spectrograms. Examples of such mixed-resolution distributions include the auditory, super-resolution, and cone-kernel spectrograms [11][12]. Towards this end, we develop in this paper signal models for the counterpart *wideband* spectrogram in the context of the GCT (WGCT) (Figure 1), thereby providing a more complete interpretation of speech signal behavior in both the GCT framework and potentially other 2-D processing schemes and time-frequency distributions.

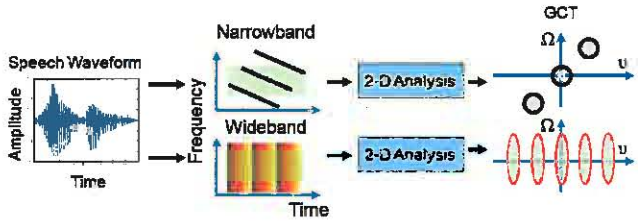

Figure 1. Schematic of general 2-D processing framework with short-time analysis followed by localized 2-D analysis for narrow (top) and wideband (bottom) representations.

In our development, we show that the WGCT is distinct from the NGCT in interpretation, thereby motivating a novel *taxonomy* of speech signal behavior in 2-D processing of speech. We also show that the WGCT can be used in speech signal processing via sinusoidal-series-based demodulation as in [10] to motivate spectrogram analysis/synthesis methods. To assess the ability of the model to represent speech content, we evaluate these methods for reconstruction of wideband spectrograms and as an example application, build on previous work in [10] in using the WGCT for co-channel speaker separation with prior pitch information. In this context, we emphasize our focus on assessing the signal models' representations of speech rather than developing a complete separation system.

This paper is organized as follows. Section II reviews the GCT framework. Section III develops a 2-D speech signal model for stationary voiced speech; Section IV describes extensions to non-stationary voiced speech while Section V discusses models for noise and onset/offset content. Section

This work is sponsored by the Department of the Air Force Rome Lab under Air Force Contract # FA8721-05-C0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government.

VI presents a taxonomy of speech signal behavior in the WGCT and NGCT. Section VII describes methods for spectrogram reconstruction and speaker separation. Sections VIII and IX present our results and conclusions, respectively.

II. Framework

Here, we review the Grating Compression Transform (GCT) framework. Consider the short-time Fourier transform (STFT) of a speech signal $y[n]$ using a window $w[n]$

$$Y(n,\omega) = \sum_{m=-\infty}^{\infty} w[m-n]y[m]e^{-j\omega m}. \quad (1)$$

In [10], we considered $w[n]$ with length (L) 2~3 times the pitch period P of voiced speech present in $y[n]$, resulting in a narrowband spectrogram. This window choice leads to *harmonic line structure oriented across frequency*. For local time-frequency regions of $|Y(n,\omega)|$

$$|Y(n,\omega)|_{local} \approx w[n,\omega]H(n,\omega)E(n,\omega) \quad (2)$$

where $w[n,\omega]$ is the 2-D window, $H(n,\omega)$ is the vocal tract formant *envelope*, and $E(n,\omega)$ is a 2-D sinusoidal-series *carrier* dependent on pitch and pitch dynamic content. In the GCT domain, the model results in *distribution* of the envelope (Figure 1). Similar behavior was argued for unvoiced speech and onset/offset content.

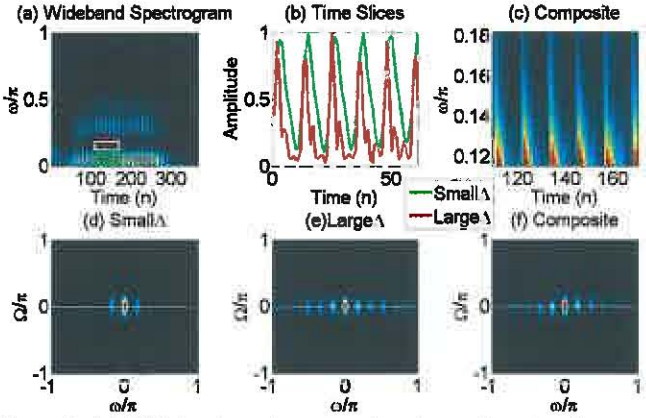

Figure 2. (a) Wideband spectrogram of real speech male utterance "needs" illustrating analysis near the first formant ($\omega \approx 0.05$) (b) small Δ (red), large Δ (green) and an (c) "edge" case (white); (d – f) WGCT representation of three regions; note off-axis terms in (e); (d – e) computed for regions including time slices in (b); see discussion of simulations for WGCT computation details.

This paper considers $w[n]$ with $L < P$, such that $w[n]$ analyzes $y[n]$ within a single period P voiced speech [12]. This window choice leads to *harmonic structure oriented across time* in a *wideband spectrogram*. A model for wideband spectrograms of voiced speech is proposed in [12]

$$|Y(n,\omega)| = E[n]\tilde{H}(\omega) \quad (3)$$

where $E[n]$ is a time-dependent term "energy" term and $\tilde{H}(\omega)$ is a "smoothed" version of the true formant envelope. Figure 2 shows analysis of several local time-frequency regions of a wideband spectrogram computed for voiced speech. We observe distinct behaviors in each region and their corresponding WGCTs based on the proximity to the first formant. Subsequently, we argue for a set of models with general form of (3) to characterize these behaviors.

III. Stationary Voiced Speech Modeling

A. *Single-Formant Modeling*

Consider a simple model of speech in which an impulse train

$$p[n] = \sum_{k=0}^{N_k} \delta[n-kP] \quad (4)$$

with periodicity P and N_k terms excites a single formant modeled as a decaying sinusoid (and Fourier transform)

$$h[n] = \xi_f e^{-\alpha_f n}\cos(\omega_f n)\,u[n]. \quad (5)$$

$$H(\omega) = \frac{0.5\xi_f}{\alpha_f + e^{j(\omega-\omega_f)}} + \frac{0.5\xi_f}{\alpha_f + e^{j(\omega+\omega_f)}}. \quad (6)$$

ξ_f, α_f, and ω_f are the amplitude, decay rate (corresponding to formant bandwidth), and formant frequency, respectively. We analyze the resulting signal

$$y[n] = \sum_{k=0}^{N_k} h[n-kP] \quad (7)$$

using the short-time Fourier transform (STFT) with $w[n]$ of length $L < P$ to satisfy the wideband constraint.

Consider the *filterbank* view of the STFT such that at an analysis frequency $\omega = \omega_f + \Delta$ [12],

$$Y(n,\omega) = (y[n]e^{-j\omega n}) *_n w[n] \quad (8)$$

$$Y(n,\omega) = \left(\sum_{k=0}^{N_k} h[n-kP]\, e^{-j(\omega_f+\Delta)n}\right) *_n w[n] \quad (9)$$

By linearity of convolution, a single term in the summation is

$$Y(n,\omega;k) = \left(h[n-kP]e^{-j(\omega_f+\Delta)n}\right) *_n w[n] \quad (10)$$

with corresponding Fourier transform

$$Y(\omega',\omega;k) = e^{-jkP(\omega'-\omega_f-\Delta)}W(\omega') \left(\frac{0.5\xi_f}{\alpha_f + e^{j(\omega'-\Delta)}} + \frac{0.5\xi_f}{\alpha_f + e^{j(\omega'-2\omega_f-\Delta)}}\right). \quad (11)$$

n maps to ω' through the Fourier transform and is distinct from ω. Since $W(\omega')$ is concentrated near $\omega' = 0$ and nearly zero far away from $\omega' = 0$ origin (i.e., at $\omega' = 2\omega_f + \Delta$),

$$Y(\omega',\omega;k) \approx e^{jkP(\omega_f+\Delta)}e^{-j\omega' kP}W(\omega')\frac{0.5\xi_f}{\alpha + e^{j(\omega'-\Delta)}}. \quad (12)$$

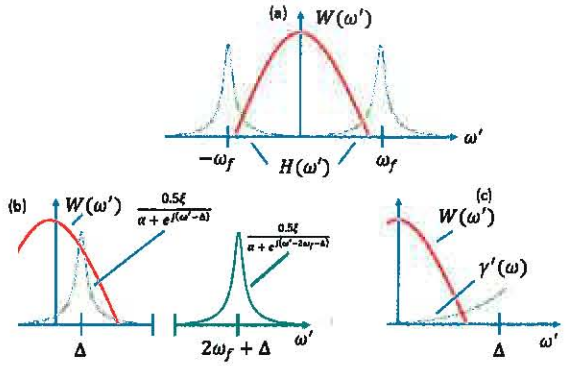

Figure 3. (a) Fourier transform of impulse response (green) and window (red); (b) small Δ case with majority of demodulated formant near origin is within window filter; modulated formant at $\omega' = 2\omega_f + \Delta$ excluded by window filter; (c) large Δ case with tail of formant content within bandwidth of window filter; $\omega' = 2\omega_f + \Delta$ component not shown.

We consider two limiting conditions of "small" or "large" values of Δ and derive *modulation* representations in both cases (Figure 3).

Small Δ: Applying the inverse Fourier transform to (12),

$$Y(n,\omega;k) \approx \left(0.5\xi_f e^{jkP(\omega_f+\Delta)}\right) w[n] *_n \left(e^{-\alpha_f(n-kP)}u[n-kP]e^{j\Delta(n-kP)}\right). \quad (13)$$

For *small* Δ, $e^{-jn\Delta}$ fluctuates slowly in time, and we therefore approximate it as $e^{-jn\Delta} \approx cos(\Delta) + jsin(\Delta)$. Furthermore, we assume that $\cos(\Delta)$ dominates $e^{-jn\Delta}$ for small Δ and $jsin(\Delta) \approx 0$. We then have

$$Y(n,\omega;k) \approx \gamma(\omega)e^{jkP(\omega_f+\Delta)}\epsilon[n-kP] \quad (14)$$

$$\gamma(\omega) = 0.5\xi_f cos\,(\omega-\omega_f) \quad (15)$$

$$\epsilon[n-kP] = w[n] *_n e^{-\alpha_f(n-kP)} u[n-kP]. \quad (16)$$

In (15), we have rewritten $\cos(\Delta)$ as $\cos(\omega-\omega_f)$ since $\Delta = \omega - \omega_f$. *Note that if $\Delta = 0$, (14) holds with equality with $\gamma(\omega) = 0.5\xi_f$.* Returning to the summation over k, we obtain

$$Y(n,\omega) \approx \sum_{k=0}^{N_k} e^{jkP(\omega_f+\Delta)}\gamma(\omega)\epsilon[n-kP]. \quad (17)$$

If $\epsilon[n-kP]$ decays to zero within each period, the *magnitude* of the sum may be approximated as the sum of the magnitudes, i.e.,

$$|Y(n,\omega)|_{local} \approx w[n,\omega]\gamma(\omega)E_d[n] \quad (18)$$

$$E_d[n] = \sum_{k=0}^{N_k}\epsilon[n-kP] = \sum_{l=0}^{N_s}\beta_l\cos\left(\frac{2\pi l}{P}n+\psi_l\right) \quad (19)$$

where $\gamma(\omega)$ is assumed to be non-negative for "small"-Δ e.g., $0 \le |\Delta| \ll \frac{\pi}{2}$, and we have rewritten $E_d[n]$ as a sinusoidal series expansion. Here, we have also introduced a 2-D analysis window $w[n,\omega]$ to emphasize analysis in a *local* time-frequency region of the wideband spectrogram.

Our derivation argues for a modulation model as in (3) with a sinusoidal series carrier $E_d[n]$ representing source periodicity and formant bandwidth/decay rate (Figure 2b) and envelope $\gamma(\omega)$ representing frequency-dependent scaling of the formant peak in the spectral domain. It can be shown from an alternative Fourier transform view of the STFT that one interpretation of this scaling is smoothing of the true formant spectrum with the Fourier transform of the window,

$$\gamma(\omega) \approx \widetilde{H}(\omega) = |W(\omega) *_\omega H(\omega)|. \quad (20)$$

We refer the reader to Appendix I for a discussion of this derivation and illustrate subsequently through simulations its limitations. Note that if the bandwidth of $W(\omega)$ is substantially greater than that of $H(\omega)$, then the bandwidth of $\widetilde{H}(\omega)$ effectively becomes that of the window.

Since $E_d[n]$ and $\gamma(\omega)$ are separable in (18), its 2-D Fourier transform (i.e., the WGCT) is

$$Y(\upsilon,\Omega) = W(\upsilon,\Omega) *_{\upsilon,\Omega} \left[\eta(\Omega)\left(K\delta(\upsilon) + \sum_{l=1}^{N_s} 0.5\beta_l e^{\mp j\psi_l\upsilon}\delta\left(\upsilon \pm \frac{2\pi l}{P}\right)\right)\right] \quad (21)$$

where n and ω map to υ and Ω, respectively, and $\eta(\Omega)$ ($W(\upsilon,\Omega)$) is the Fourier transform of $\widetilde{H}(\omega)$ ($w[n,\omega]$). $\eta(\Omega)$ is the WGCT representation of the smoothed formant envelope in a *local* time-frequency region. Copies of $\eta(\Omega)$ are weighted by β_l coefficients (representing the bandwidth of the formant) along the υ-axis at multiples of $\frac{2\pi}{P}$ (representing the source periodicity). This product is further smoothed in υ and Ω with the Fourier transform of the 2-D analysis window. Note that formant bandwidth along the ω-axis "lost" due to smoothing by the short-time analysis window is "recovered" in *time* and represented in the carrier.

The present and subsequent formulation motivates modulation/demodulation framework for speech signal processing similar to [10]. Since copies of $\eta(\Omega)$ are distributed in the WGCT space via the carriers, they may be *demodulated* to reconstruct the $\eta(\Omega)$ term at the WGCT origin if this component is corrupted e.g. from an interfering signal.

Large Δ: For large Δ, the approximation in (14) does not hold. $\omega = \omega_f + \Delta$ is "far away" from ω_f, and we alternatively assume that the frequency response of the formant is approximately a single complex value $\gamma'(\omega)$ (i.e., a flat spectrum) (Figure 3). The frequency domain interpretation of this from (11) is

$$Y(\omega',\omega;k) = \gamma'(\omega)(e^{-jkP(\omega'+\omega)}W(\omega'). \quad (22)$$

Inverting (22) and invoking the summation as in (17),

$$Y(n,\omega;k) = \sum_k^{N_k}\gamma'(\omega)e^{-jkp\omega}\delta(n-kP) *_n w[n] = \sum_k^{N_k}\gamma'(\omega)e^{-jkp\omega}w[n-kP] \quad (23)$$

Since $L < P$, the summed terms do not overlap in time. In a local-time frequency region analyzed with a 2-D window $w[n,\omega]$, the magnitude of the sum can be rewritten as a sum of magnitudes, i.e.,

$$|Y(n,\omega)|_{local} = w[n,\omega]|\gamma'(\omega)|E_w[n] \quad (24)$$
$$E_w[n] = \sum_k^{N_k} w[n-kP] \quad (25)$$

resulting again in a *modulation* model of the spectrogram with a source periodicity-dependent carrier $E_w[n]$ and an envelope $|\gamma'(\omega)|$ which we again interpret as $\tilde{H}(\omega)$ from (20). An analogous WGCT representation is (Figure 4)

$$Y(\upsilon,\Omega) = W(\upsilon,\Omega) *_{\upsilon,\Omega} \left[\eta'(\Omega)\begin{pmatrix} K\delta(\upsilon) + \\ \sum_{l=1}^{N_s} 0.5\beta'_l e^{\mp j\psi'_l \upsilon}\delta\left(\upsilon \pm \frac{2\pi l}{P}\right)\end{pmatrix}\right] \quad (26)$$

where $\eta'(\Omega)$ is the Fourier transform of $|\gamma'(\omega)|$ and β'_l and ψ' parameters of the sinusoidal series representation of $E_w[n]$. While the WGCT domain contains copies of $\eta'(\Omega)$ reflecting smoothed formant structure in local time-frequency regions as in the small Δ case, carrier positions and corresponding gain terms reflect source periodicity only.

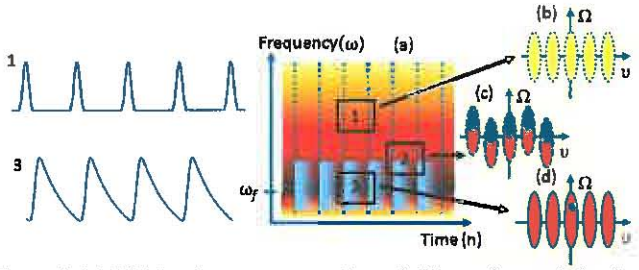

Figure 4. (a) Wideband spectrogram schematic illustrating analysis of a single formant in distinct frequency regions (1) large Δ, (2) small Δ, (3), "in between" case; periodicity and bandwidth-dependent carrier (blue, shaded), periodicity-dependent carrier (dotted lines) and composite carrier; (b-d) WGCT of regions 1 – 3 with distinct modulated envelopes delineated: small Δ – red, large Δ – yellow, "in between" – graded.

Composite Carrier: Our discussion thus far has described for limiting cases of Δ modulation models in time-frequency regions of wideband spectrograms. To account for values of Δ "in between", we propose a "composite" carrier

$$E_c[n,\omega] = E_d[n]R[\omega] + E_w[n]R[\omega-\omega_0] \quad (27)$$
$$R[\omega] = \begin{matrix}1, 0<\omega<M \\ 0, otherwise\end{matrix} \quad (28)$$

where M ranges from 0 to the full length M_{full} region in frequency, and ω_0 is a shift in frequency. A similar composite carrier can be obtained by interchanging $E_w[n]$ and $E_d[n]$. $E_c[n,\omega]$ may be modulated by $\tilde{H}(\omega)$ to invoke a modulation interpretation as in the limiting cases. A *generalized* modulation model in local time-frequency regions is

$$|Y(n,\omega)|_{local} = w[n,\omega]\tilde{H}(\omega) \quad (29)$$
$$(E_d[n]R[\omega] + E_w[n]R[\omega-\omega_0]).$$

The 2-D Fourier transform (WGCT) of (29) is (Figure 4)

$$Y(\upsilon,\Omega) = W(\upsilon,\Omega) *_{\upsilon,\Omega} \sum_{i\epsilon(d,w)} \eta_{R,i}(\Omega)\begin{pmatrix} K_{R,i}\delta(\upsilon) + \\ \sum_{l=1}^{N_s} 0.5\beta_{l_{R,l}}\delta\left(\upsilon \pm \frac{2\pi l}{P}\right)\end{pmatrix} \quad (30)$$

where $\eta_{R,i}(\Omega)$ is the Fourier transform of $\tilde{H}(\omega)R(\omega)$ $(i=d)$ and $\tilde{H}(\omega)R(\omega-\omega_0)$ $(i=w)$. $K_{R,i}$ and $\beta_{l_{R,i}}$ is a complex coefficient corresponding to the sinusoidal series of the two carrier types. The WGCT contains a scaled *sum* of $\eta_{R,i}(\Omega)$ terms at the origin and carrier locations. If the bandwidth of $\eta_{R,i}(\Omega)$ υ_η are such that $0.5\upsilon_\eta < \frac{2\pi}{P}$, then their modulated copies will occupy distinct regions along the υ-axis (Figure 2). Note that this model does not impose constraints on the bandwidth along the Ω-axis.

The WGCT also invokes a mapping of pitch f_0 information

$$\upsilon_0 = f_0\frac{2\pi}{f_s} \quad (31)$$

where f_s is the sampling frequency of the waveform. If the time width of the local time-frequency region is be 2~3 times the pitch period [12], the resulting the WGCT exhibits distinct copies of the envelope at multiples of $\frac{2\pi k}{P}$; f_0 is *inversely* related to the number of terms in the WGCT.

B. *Multiple Formants*

For multiple formants, we generalize (11) as the summation

$$Y(\omega',\omega;k) = W(\omega') \sum_{f=1}^{N_f} e^{jkP(\omega'+\omega_f+\Delta)}\left(\frac{0.5\xi}{\alpha_f+e^{j(\omega'-\Delta)}} + \frac{0.5\xi}{\alpha_f+e^{j(\omega'-2\omega_f-\Delta)}}\right). \quad (32)$$

where N_f is the number of formants. Assuming that the ω_f are well-separated in frequency, we approximate $Y(\omega',\omega;k)$ as being dominated by a *single* formant in local frequency regions. Consequently, identical arguments can be applied as in the previous sections to arrive at modulation models for individual formants. This invokes a sum-of-magnitudes approximation for the magnitude

$$|Y(n,\omega)|_{local} \approx w[n,\omega]\sum_{f=1}^{N_f} E_c[n,\omega;f]\tilde{H}_f(\omega) \quad (33)$$

where $E_c[n,\omega;f]$ and $\tilde{H}_f(\omega)$ are formant specific. This model interprets local regions of the wideband spectrogram magnitude as a *sum of modulation products*. The WGCT $Y(\upsilon,\Omega)$ is a summation of terms as in (33) for each formant

$$Y(\upsilon,\Omega) = W(\upsilon,\Omega) *_{\upsilon,\Omega}$$

$$\sum_{f=1}^{N_f}\sum_{l\epsilon(d,w)}\eta_{R,l}(\Omega;f)\begin{pmatrix}K_{R,i,f}\delta(v)+\\ \sum_{l=1}^{N_s}0.5\beta_{l_{R,i,f}}\delta\left(v\pm\frac{2\pi l}{P}\right)\end{pmatrix}$$

where $\eta_{R,i}(\Omega;f)$, $K_{R,i,f}$, and $\beta_{l_{R,i,f}}$ are now formant-dependent versions of those in (30). We expect this approximation to be best for frequency regions near a formant peak, e.g., $\omega_f-\Delta<\omega<\omega_f+\Delta$, analogous to the single-formant case. Furthermore, at frequency regions far away from formant frequencies, the summation implies dominance by a "large Δ" model (24) corresponding to a single formant. Nonetheless, if formants interact within a local frequency region, the model can be expected to less accurate. In our subsequent analyses, we show the effects of such interactions.

C. *Simulations*

Single Formant: Herein we illustrate properties of the *carrier* models proposed for the previously described small and large Δ conditions. We synthesize a decaying sinusoid $h'[n]$ with $\omega_f=0.1\pi$ corresponding to a periodicity of 20 samples, $\xi=1$, and $\alpha_f=0.01$ (5); $h'[n]$ is excited with a pure impulse train $p'[n]$ with periodicity $P=77$ to generate $y'[n]$. Signals are synthesized at 16 kHz with resulting pitch (formant) frequency of 210 Hz (800 Hz). Wideband spectrograms are computed using a Hamming window $w[n]$ with length $L=40\equiv 2.5$-ms Hamming; to account for an extremal case of a 350-Hz pitch, L can be chosen in general to be less than $\sim\frac{1}{350}=2.9$-ms. A single-sample frame rate and 2048-point discrete Fourier transform (DFT) is applied to both $y'[n]$ and $p'[n]$ to obtain $|Y'(n,\omega)|$ and $|P'(n,\omega)|$. WGCT analysis was performed using region sizes of 37.5-ms by 500 Hz extracted with a 2-D Hamming window followed by a 512 by 512-point 2-D DFT. Analogous to the choice of L, 2~3 times the lowest pitch period of 60 Hz constrains the time width to ~33 to 50 ms. We refer the reader to our subsequent discussion to motivate the choice of frequency widths.

For "small-Δ", we extract time slices from $|Y(n,\omega)|$ at $\omega=\omega_f$ and $\omega=\omega_f+\Delta$ with $\Delta=0.0313\pi$ (corresponding to 250 Hz). $\omega=\omega_f$ ($\Delta=0$) represents the *idealized carrier* in the modulation model as discussed in (14); $\Delta=0.0313\pi$ represents a "small-Δ" condition. Time slices are normalized to have unity amplitude and shown in Figure 5b. We plot absolute differences between the slices and compute the root-mean-squared error (RMSE) across time. Consistent with the model, both time slices resemble decaying exponentials smoothed by the window with the $\Delta=0.0313\pi$ case having RMSE of ~0.09 relative to the $\Delta=0$ case. This discrepancy is presumably due to phase effects ignored in modeling.

For "large-Δ", Figure 5c shows a time slice extracted at $\omega=0.5\pi$ (i.e., "far away" from ω_f). We also plot a time slice $|P'(n,\omega)|$ corresponding to periodically summed windows, i.e., the idealized carrier $E_w[n]$. The $\omega=0.5\pi$ time slice closely matches $E_w[n]$ with RMSE of ~0.05.

In a second set of simulations, we explore properties of the smoothed formant interpretation of the envelope term of the modulation model (20). We replicate a time slice $|Y(n,\omega=\omega_f)|$ across all frequencies to generate a 2-D carrier $E'(n,\omega)$ and compute the time average of all spectral slices in $|Y(n,\omega)|$ and replicate this across time to obtain a reference estimate of the smoothed envelope term $\bar{H}(n,\omega)\approx H_r(n,\omega)$. Subsequently, we compute

$$E'_e(n,\omega)=\frac{|Y'(n,\omega)|}{H_r(n,\omega)},\ H_e(n,\omega)=\frac{|Y'(n,\omega)|}{E'(n,\omega)}. \quad (34)$$

Figure 6 shows $E'_e(n,\omega)$ and time slices corresponding to the decaying and window-based carriers in regions near and far from the ω_f respectively, as can be expected since $H_r(n,\omega)$ varies with frequency only. In addition, $H_e(n,\omega)$ is reasonably matched to $H_r(n,\omega)$ in frequency regions near ω_f though not for ω away from ω_f. This is consistent with our use of the exponential decaying carrier in computing $H_e(n,\omega)$. In addition, $H_e(n,\omega)$ exhibits temporal fluctuations in energy at ω_f. This effect reflects the fact that the assumed envelope $H_r(n,\omega)$ best matches in time regions away from excitation impulses (see Appendix I). Quantitatively, normalized spectral slices of $H_e(n,\omega)$ exhibit an RMSE relative to $H_r(n,\omega)$ up to ~0.09.

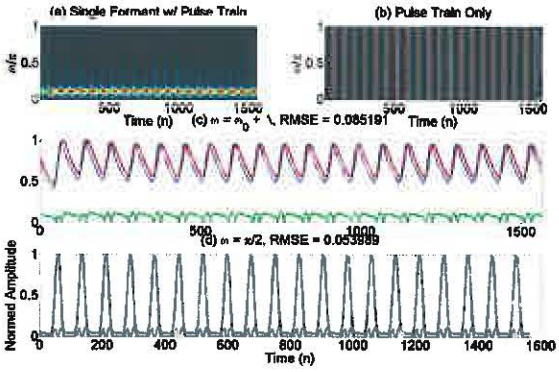

Figure 5. Wideband spectrogram of (a) decaying sinusoid excited with a pure impulse train and (b) pure impulse train; note that a time slice of (b) corresponds to periodically summed copies of the short-time analysis window; (c) time slice of (a) located at the formant peak (red) and for a small Δ value away from the peak (blue); absolute difference (green) between the two curves; (d) as in (c) but for the idealized pure impulse train time slice (red) and actual time slice located "far away" from the formant peak (blue).

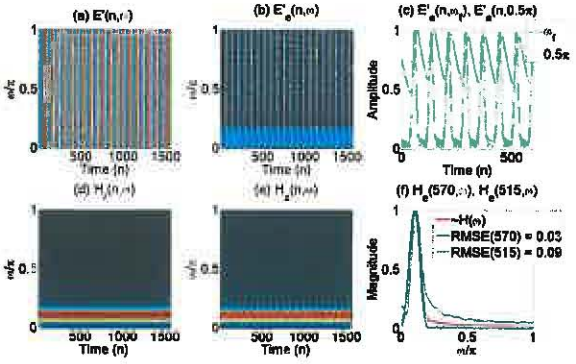

Figure 6. (a) $Y(n,\omega_f)$; (b) $E'_e(n,\omega)$; (c) time slices of (b); (d) $H_r(n,\omega)$; (e) $H_e(n,\omega)$; (f) spectral slices of (e) and (f); RMSEs in (f) computed between normalized spectral slices of (e) and the reference estimate in (d).

In a final set of simulations, we assess model properties in frequency regions in between the limiting cases of "small"

and "large" Δ. Figure 6 shows a local region of $E'_e(n,\omega)$ (34) centered at $\omega = 0.23\pi$ in which two carriers appear to interact within the same local region. The corresponding WGCT contains components *off* the horizontal axis, violating the assumption of a strictly time-dependent carrier ($E_d[n], E_w[n]$). From (27), we set each half of the region in frequency to $E_d[n]$ and $E_w[n]$. Observe that the resulting WGCT of this signal does indeed exhibit off-axis similar to those in Figure 7a. In Figure 7c, we show the result of summing $E_w[n]$ and $E[n]$ *without* applying $R[\omega]$; the resulting WGCT does *not* exhibit off-axis terms, indicating that the displacement effects of $R[\omega]$ corresponding to phase terms in the WGCT are crucial in modeling this behavior. This can understood from (30) by noting that the Fourier transform of $E_c[n,\omega]$ has the same form but with $\eta_{R,i}(\Omega)$ replaced by the Fourier transforms of $R[\omega]$ and $R[\omega-\omega_0]$, thereby invoking dependence along Ω in the WGCT domain.

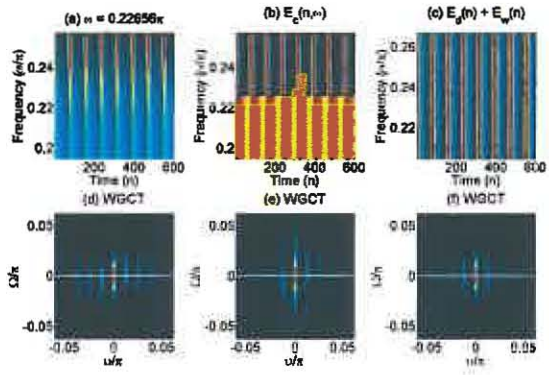

Figure 7. (a) Local region from $E'_e(n,\omega)$ centered at $\sim 0.23\pi$; (b) "composite" carrier; (c) carrier obtained from direct summation; (d-f) WGCT of (a-c), respectively with $\Omega = 0$ (line) denoted.

Multiple Formants: Herein we explore properties of the sum of modulation products model (33) for multiple formants. In addition, we implicitly investigate effects of downsampling the spectrogram, as is typically done in implementation, on the model. Furthermore, we motivate a choice of region size in WGCT analysis along the frequency dimension.

A synthetic vowel generated using a pure impulse train $p[n]$ with a 250-Hz pitch is filtered with a stationary formant structure with frequencies (bandwidths) 669, 2349, 2972, 3500 Hz (65, 90, 156, 200 Hz) to generate $y[n]$ (i.e., a female /ae/ vowel, [13]). Spectrograms are computed as in the previous section though a frame rate of 10 samples (i.e., $\frac{L}{4}$). In addition, we apply a high-pass filter to the spectrogram and aim to recover localized regions using demodulation with bootstrapping as alluded to in Section III.A. For each point along the ω-axis, we extract a region of the filtered spectrogram of time length 37.5 ms and vary the frequency width to obtain local regions (Figure 8d). Using demodulation, we obtain an estimate of the original local region; we refer the reader to VII.A for details of the method and focus here on the results. We compute the root-mean-squared error (RMSE) between the estimate and original 2-D region extracted after both are scaled to have maximum value of unity for comparison purposes.

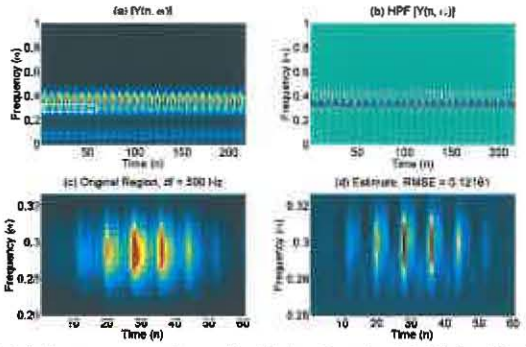

Figure 8. (a) Spectrogram of vowel with local region highlighted (white); (b) high-pass filtered version of (a) for use in reconstruction; (c) original local region; (d) estimate of (d) using demodulation.

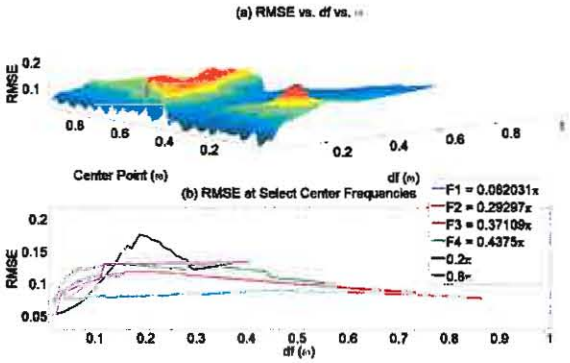

Figure 9 (a) RMSE as a function of frequency widths and frequency points analyzed; (b) RMSE for frequency center points corresponding to formant frequencies.

Figure 9a shows results across all frequency center points and widths (*df*). Figure 9b shows results of analysis at select center frequencies. Despite the presence of multiple formants, RMSEs for reconstructions centered at the formant frequencies do not exceed ~0.15 for frequency widths ranging from zero to 0.1π corresponding to ~800 Hz and result in reasonable estimates of the original region. RMSE values generally increase up to a local maximum for larger widths followed by a modest decrease. At frequency regions "far away" from formant peaks, e.g., at $\omega = 0.8\pi$, reconstructions also follow this trend though substantially less growth in RMSE beyond frequency widths of 0.1π; this is due to the absence of interacting formant structure in these frequency regions. Conversely, at $\omega = 0.2\pi$, the slope of the RMSE is sharper than for the individual and $\omega = 0.8\pi$ case, reflecting effects of formant interactions (here, F1 and F2).

IV. Extensions to Non-stationary Voiced Speech

A. *Dynamic Formants*

Modeling: As discussed in Appendix I, a Fourier transform view of the wideband spectrograms argues for a similar modulation model to (29) that includes formant dynamics. In the time-frequency space, we view dynamic formant content as a *rotated* rectangle in the time-frequency space such that the 2-D Fourier transform is the rotation of the 2-D Fourier transform of a rectangle from image processing principles (Figure 18) [10]. While the derived model is posed under relatively restrictive conditions in relation to time segments

away from excitation impulse onsets, herein we illustrate with a simple example that the model can nonetheless provide a reasonable interpretation of dynamic formants.

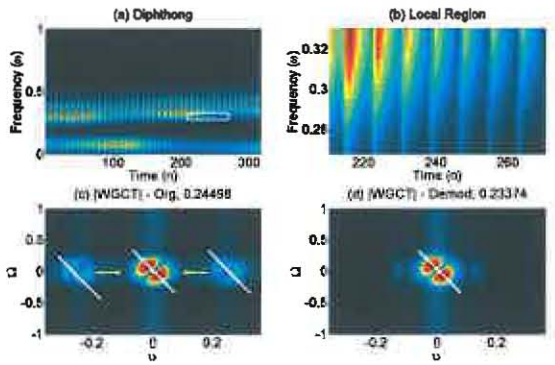

Figure 10. (a) Wideband spectrogram of diphthong with local region (white); (b) local region of (a); (c) GCT of (b) with *rotated* (white line) envelope structure near origin; arrows denote demodulation of carrier terms down to DC; (d) WGCT of *demodulated* version of (c) with comparable rotated components to match that in (c). In (c) and (d), DC value is removed for illustrative purposes; in (d), display limited to near-DC region due to presence of cross terms in demodulation.

Dynamic Formant Model Simulations: We synthesize a 200-ms diphthong with start-to-end formant frequencies (bandwidths) of 669, 2349, 2972, 4000 Hz (65, 90, 156, 200 Hz) to 437 2761, 3372, 4000 Hz (38, 66, 171, 200 Hz). The source signal is a pure impulse train with 200-Hz pitch. The wideband spectrogram and WGCTs are computed as in the previous section.

Figure 10a-b shows a local region near the increasing second formant. Figure 10c shows the corresponding WGCT near the first carrier position; for display purposes, DC values at both the origin and carriers have been removed. At these locations, we observe *rotated* components corresponding to the local envelope structure present in Figure 10b; the rotation of these components can be quantified by measuring the angle of the near-DC peaks relative to the Ω-axis of ~0.24 radians.

As noted in Section III.A, demodulation of envelope content from carrier positions may be used to recover near-DC terms in the WGCT. Figure 10d shows an example of demodulating the carrier components in Figure 10c to DC. Since in reconstruction we further remove any resulting cross terms by low-pass filtering (see Section VII), we restrict our display to the near-DC regions here. A set of *rotated* components are obtained at DC with angle ~0.23 radians to match those at DC in Figure 10c. These results are consistent with a generalized 2-D envelope $\tilde{H}(n,\omega)$ as argued in Appendix I in relation to the modulation model.

B. *Time-varying Pitch*

***Model*:** Time-varying models of pitch have been explored by a number of researchers such as in [14]. In the short-time spectrum, the behavior of time-varying impulse has been described qualitatively as "blurring" (i.e., widening) of harmonic peaks near the "average" pitch; this effect may be interpreted as multiple peaks in the spectrum corresponding to a Bessell function expansion [12]. In our present development, we impose the constraint that local time-frequency regions have time widths such that pitch values are approximately constant. Subsequently, we quantitatively assess the effect this has on a range of pitch variations.

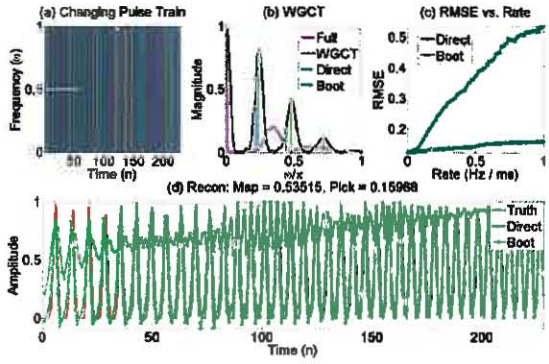

Figure 11. (a) Wideband spectrogram of changing pitch with time segment (37.5 ms) denoted (white); (b) WGCT of full time slice (maroon) and time segment of (a) (black); peaks obtained in direct mapping (blue) and bootstrapping (green); (c) RMSE of reconstructions using direct versus bootstrapping methods; (d) reconstruction of 1 Hz/ms case with truth (red), direct (blue), and bootstrapping (green) denoted.

Time-varying Pitch Simulations: We synthesis impulse trains of duration 200 ms with linearly increasing pitch (varied fro m0 to 1 Hz/ms) with starting pitch value of 175 Hz. In analysis, we compute the wideband spectrogram and attempt to resynthesize a full time slice from time segments of size 37.5 ms using 1-D WGCT analysis (Figure 11). We extract "peak" locations in the WGCT to resynthesize a sinusoidal series. Peak locations are determined using either 1) the direct pitch information mapping of (31) or 2) bootstrapping of the peak locations (Figure 11b). In the former method, the pitch value defined at the center of the time segment is used; in the latter, the mapped locations are reassigned using a 1-D multi-peak picker applied to the WGCT (see Section V and [10]).

The resulting WGCT shows that the direct mapping can result in "peak" locations that appear harmonically realted but deviate from the actual WGCT peaks (Figure 11b). As an extremal example of the variation in peak location with time-varying pitch, Figure 11b shows a GCT computed for the *full* time slice; we observe two peaks with substantially widened bandwidths consistent with the previously described Bessel-like behavior. We compute the root-mean-squared error (RMSE) between normalized estimates and true time slices. Figure 11c shows that RMSE increases dramatically using the direct method for rates > ~0.1 Hz/ms in contrast to the bootstrapping technique. At a rate of 1 Hz/ms, bootstrapping (RMSE = ~0.16) maintains the aperiodicity of the signal while the direct mapping (RMSE = ~0.54) deviates substantially motivates a *bootstrapping* approach to obtain carrier locations that may not correspond exactly to the pitch mapping of (31).

V. Noise and Onsets/Offsets Models

A. *Noise*

Model: We consider now modeling of noisy signals (e.g., fricatives) in the WGCT. The analytical form of the WGCT model of noise is identical to that presented for the Narrowband GCT (NGCT), and we refer the reader to [10] for

more details while focusing on empirical behavior of noise in the WGCT in this section.

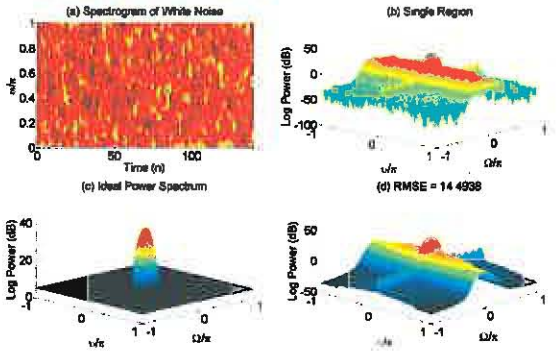

Figure 12. (a) Wideband spectrogram of white noise (log-scale); (b) WGCT of a single region; (c) ideal average power spectrum; (d) estimated average power spectrum.

A zero-mean independent and identically distributed (i.i.d.) Gaussian process $\boldsymbol{w}[t]$ with standard deviation σ can be analyzed with a wideband short-time Fourier transform magnitude $\boldsymbol{w}[n,\omega]$. We model $\boldsymbol{w}[n,\omega]$ as arising from a 2-D random process under assumptions of i.i.d. time-frequency units with Rayleigh distribution. The (average) 2-D power spectrum of $\boldsymbol{w}[n,\omega]$ is then [10],

$$S_{ww,GCT}(\upsilon,\Omega) = \left[\frac{4-\pi}{2}\sigma^2 + \frac{\pi}{2}\sigma^2\delta(\upsilon,\Omega)\right] *_{\upsilon,\Omega} |W(\upsilon,\Omega)|^2 \quad (35)$$

$$= \frac{\pi}{2}\sigma^2|W(\upsilon,\Omega)|^2 + \frac{4-\pi}{2}\sigma^2\rho$$

$$\rho = \iint_{(-\pi,-\pi)}^{(\pi,\pi)} |W(\upsilon,\Omega)|^2 d\upsilon d\,\Omega \quad (36)$$

where $W(\upsilon,\Omega)$ is the 2-D Fourier transform of the 2-D window used to extract localized regions of $\boldsymbol{w}[n,\omega]$.

To obtain an instantaneous model, we invoked in [10] the Karhunen-Loeve expansion under the assumption of distinct frequency bands of the filterbank view of the 2-D Fourier transform [12]. Specifically, a sum of arbitrary sinusoids on a DC pedestal was viewed as the *carrier* component in the modulation model of (29) (and corresponding WGCT)

$$|Y(n,\omega)| = w[n,\omega]\tilde{H}[n,\omega]E[n,\omega] \quad (37)$$

$$E(n,\omega) = K + \sum_{k=1}^{N_c} \alpha_k \cos(\phi_k[n,\omega]) \quad (38)$$

$$\phi_k[n,\omega] = \Omega_k(n\cos\theta + \omega\sin\theta) + \varphi_k \quad (39)$$

$$Y(\upsilon,\Omega) = W(\upsilon,\Omega) *_{\upsilon,\Omega} \left(\sum_{k=1}^{N_c} K\eta(\upsilon,\Omega) + 0.5\alpha_k\eta(\upsilon \pm \Omega_k\cos\theta, \Omega \pm \Omega_k\sin\theta)\right) \quad (40)$$

with N_c as the number of carriers, Ω_k is the spatial frequency of the 2-D sinusoid, θ its orientation, and φ_k its phase term. Here, we have where we have allowed for a 2-D envelope $\tilde{H}[n,\omega]$ as in the time-varying formant condition. As in the voiced case, this model argues for a *distribution* of envelope content in the WGCT space at carrier locations (Figure 15f).

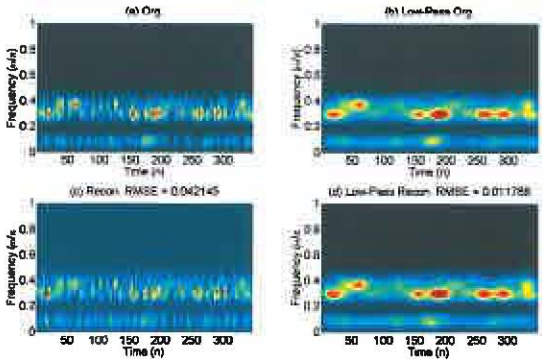

Figure 13. (a) Original spectrogram of noise-excited vowel; (b) low-pass filtered version of (a) resulting in envelope term; (c) reconstruction after high-pass filtering and demodulation; RMSE computed between (c) and (a); (d) low-pass filtered version of (c) indicating recovery of low-pass envelope term in (b); RMSE computed between (d) and (b).

Simulations: Herein we compute the empirical 2-D power spectra of white noise in wideband spectrograms for comparison to the proposed model. Figure 12a shows a wideband spectrogram computed for $\boldsymbol{w}[t]$ with $\sigma = 1$. WGCT analysis was performed using the parameters described previously for vowels. Figure 12d illustrates the power spectrum obtained from averaging all regions analyzed. While the model captures the dominance of the near-DC region of the WGCT, it fails to capture substantial 2-D spectral shaping effects. Figure 12b shows WGCT analysis results for a single region, consistent with the averaged spectrum in Figure 12d. The estimated spectrum has a substantial component along the υ-axis due to correlation across the frequency axis (ω) in the wideband spectrogram such that we observe vertical striations (Figure 12a) across time. Specifically, the short-time spectrum is substantially smeared across ω due to the relatively *short* length of the window (and therefore *wide* bandwidth in the spectrum). This behavior is the "dual" of the narrowband GCT that exhibited components along the Ω-axis due to *temporal* correlation effects of processing noise.

In a second set of simulations, we aim to assess the extent to which (37) can represent noise speech. We compute the wideband spectrogram of a vowel with formant structure as in the previous sections but excited with Gaussian white noise. Next, we adopt the framework of the the simulations for multiple formants in *removing* DC components in the WGCT with the aim of approximately reconstructing them through demodulation (Section III.A, Section V). Figure 13 shows reconstruction results and low-pass filtering of the original and reconstruction. Observe that the reconstruction results in recovery of the low-pass envelope to match that of the original spectrogram; this is consistent with the demodulation process recovering the near-DC terms of the WGCT from its distributed copies due to the carrier.

B. *Onsets/Offsets*

Model: Similar to the noise case, herein we briefly describe onset/offset content observed in wideband spectrograms and similar to that observed for the narrowband case [10]. An isolated impulse $i[n] = \delta[n - N_0]$ located at N_0 can be modeled as a downsampled short-time analysis window $w_t[n]$ in the spectrogram domain (denoted as $I[n,\omega]$)

$$I[n,\omega] = w_n[N_0 - nN] \qquad (41)$$

where N is the frame rate of the STFT. The GCT is

$$I(\upsilon,\Omega) = W(\upsilon,\Omega) *_{\upsilon} W_n^*\left(\frac{\upsilon}{N}\right) e^{j\upsilon N_0} \qquad (42)$$

where $*_{\upsilon,\Omega}$ denotes convolution in the GCT domain and $W(\upsilon,\Omega)$ is the 2-D Fourier transform of a 2-D window $w[n,\omega]$ used to extract a localized time-frequency region. We view $I(\upsilon,\Omega)$ as an envelope term in the modulation model of (29) in the context of a carrier due to voiced (e.g., (27)) or noisy speech. As with formant envelopes, we impose a a bandlimited constraint on $I(\upsilon,\Omega)$ in the context of modulation (30). In the wideband case, $I(\upsilon,\Omega)$ will have larger bandwidth than in the narrowband case due to the sharpness of the representation in time. Specifically, wideband parameters are a 2.5-ms ($L = 40$) short-time analysis window and frame rate of 0.625 ms. This results in the Hamming window mainlobe $W_n^*\left(\frac{\upsilon}{N}\right)$ width of $\left(\frac{8\pi}{40}\right)4 = 0.8\pi$ [10]; in contrast, a 32-ms window, 1-ms frame rate in the narrowband case results in a mainlobe width of $\left(\frac{8\pi}{512}\right)25 = 0.3906\pi$.

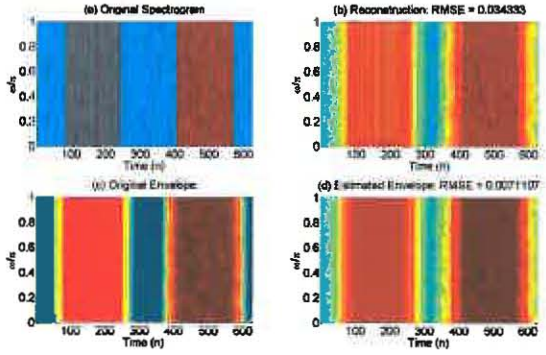

Figure 14. (a) Spectrogram of voicing and noise onset; (b) reconstruction of (a); (c) low-pass filtered version of (a) demonstrating onset/offset envelopes; as in (c) but for the reconstruction in (b); associated RMSEs computed after normalization in all cases; log spectrograms plotted to emphasize widening effects.

Simulations: Figure 14. shows results of synthesizing and reconstructing voicing and noise onset/offsets using demodulation (as done in the noise case). The reconstruction in Figure 14.b exhibits widening of the onsets as may be expected from the bandlimited nature of the analysis/synthesis method. Nonetheless, this widening is consistent with the envelope obtained in low-pass filtering the original signal in Figure 14.c and as can be shown in filtering the reconstruction in Figure 14.d.

VI. A Taxonomy of Speech Signal Behavior in the GCT

Our discussions motivated a *modulation* view of the wideband spectrogram. Specifically, in voiced regions, the wideband spectrogram can be viewed as *summation* of *modulation* components, where each component corresponds to a formant. A carrier $E_c[n,\omega]$ is dependent on source periodicity and (under certain conditions) formant bandwidth and is *modulated* by a smoothed (*single*) formant or envelope $|\tilde{H}_f(n_0,\omega)|$. Noise and onsets/offsets are viewed in this framework as carrier and envelope components, respectively.

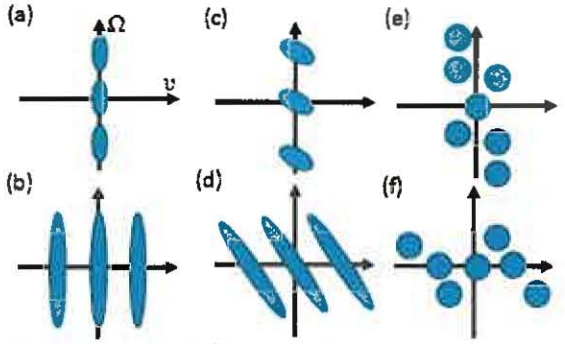

Figure 15. Narrow (top) and wideband (bottom) representations of: (a, b) stationary formant and pitch, (c, d) stationary pitch and dynamic formant, and (e, f) noise content.

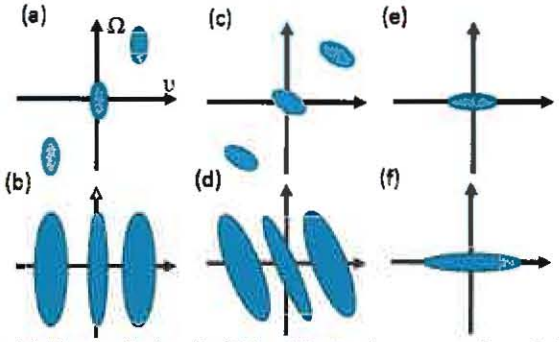

Figure 16. Narrow (top) and wideband (bottom) representations of: (a, b) dynamic pitch and stationary formant, (c, d) dynamic pitch and dynamic formant, and (e, f) onset/offset content.

This signal model has some similarity to that proposed for narrowband spectrograms [10], and in subsequent sections, we assess its ability to represent speech content using algorithms similar to those used in [10]. Nonetheless, there important distinctions exist in the form and interpretation of the two models, and in Figure 15 – Figure 16 we compare the mapping of changing/stationary, pitch/formant, and noise and onset/offset content for both representations.

For voiced speech, stationary pitch mappings in the NGCT and WGCT are "duals" of each other along the Ω (narrow) and υ (wide) axes, as schematized in Figure 15a-b. This mapping distinction is preserved even when formant dynamics are introduced (Figure 15c-d). In contrast, pitch dynamics invokes a rotation of components in the NGCT while invoking widening of the formant content along the υ-axis in the WGCT due to the presence of widened harmonic content of the carrier (Section II.C), as schematized in Figure 16a-d. An additional narrowband/wideband "duality" is observed in mapping noise to the GCT domain with components along the Ω (narrow) and υ (wide) axes (i.e., $\upsilon = 0$ and $\Omega = 0$), respectively (Figure 15e-f. Finally, the WGCT exhibits greater bandwidth of onset/offset content relative to the NGCT due to differences in short-time analysis resolution (Figure 16e-f).

Table 1 presents a taxonomy of speech signal behavior as represented in the narrowband/wideband models. We denote

$H(n,\omega)$ as the formant structure, $g(.)$ as a general function, and ω_c as the center frequency of the local region analyzed. Several distinctions include the summation of (WGCT) vs. singular modulation products (NGCT) and single (NGCT) vs. multiple carrier types (WGCT); in addition, carriers have distinct dependencies on source periodicity f_0 (NGCT, WGCT), pitch dynamics $\frac{df_0}{dt}$ (NGCT), formant bandwidth α_f (WGCT), and ω_c. "Dual" behavior exists in pitch mappings between the two GCTs; specifically, *high* pitch values results in *low* (i.e., near GCT origin) frequency components in the NGCT and *high* frequency components in the WGCT. This effect also results in the difference in number N_p of harmonic terms in the GCT as they relate to pitch. While noise is viewed as a carrier term in modulation in both representations, its localization is distinct between the two as previously noted. Finally, onsets/offsets are interpreted as envelope terms in both cases though with differences in bandwidth v_a along the v-axis.

Table 1 Comparison of signal model interpretations for narrow- and wideband-based Grating Compression Transforms.

Interpretation/GCT	Narrowband	Wideband
Local Model	$Y(n,\omega) = H(n,\omega)E(n,\omega)$	$Y(n,\omega) = \sum_{f=1}^{N_f} \tilde{H}_f(n,\omega)E_c(n,\omega)$
Envelope (vowels)	$H(n,\omega)$	$\tilde{H}_f(\omega,n;\omega_c) \approx \lvert H_f(\omega,n) * W(\omega)\rvert$
Carrier (vowels)	$E(n,\omega;f_0,\frac{df_0}{dt},\omega_c)$	$E_c(n,\omega;f,f_0,\alpha_f,\omega_c) = g(E_d(n),E_w(n),R(\omega))$
f_0 mapping	$\Omega_0 \propto \frac{1}{f_0}$; $v_0 \propto \frac{df_0}{dt}$; ω_c	$v_0 \propto \frac{1}{f_0}$
$f_0 \rightarrow N_P$	$N_P \propto f_0$	$N_P \propto \frac{1}{f_0}$
Noise	Along $v = 0$; carrier	Along $\Omega = 0$; carrier
Onsets/Offsets	$v_a = 0.39\pi$	$v_a = 0.8$

VII. Spectrogram Analysis/Synthesis and Co-channel Speaker Separation

Herein we describe approaches to test the proposed model's ability to represent speech content through spectrogram analysis/synthesis and co-channel speaker separation. As these methods are generally the same algorithmically to those in [10] and [5], we refer the reader to those works for details and focus here on the general framework and distinctions.

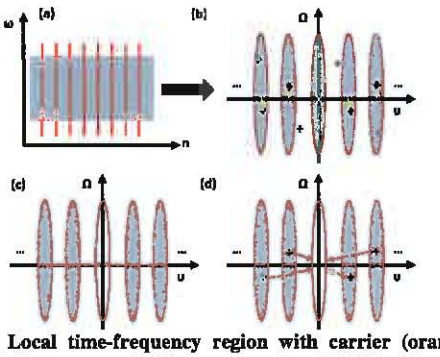

Figure 17 (a) Local time-frequency region with carrier (orange) and envelope (shaded) components; (b) corresponding WGCT with candidate peaks from peak-picking ('+') and *reassignment* of directly mapped carrier locations to candidate peaks; 'x' denotes removal of near-DC term; (c) demodulation of components located at carrier locations obtained from *direct* mapping for reconstruction; (d) as in (c) but using *reassigned* carrier locations (bootstrapping).

A. *Analysis/Synthesis*

In the proposed signal model, the WGCT domain consists of envelope content near the origin and at carrier locations (Figure 17) due to sinusoidal-series-based modulation. As a framework for reconstruction, we aim to approximately *recover* the near-DC terms in the GCT using their modulated version at carrier locations using sinusoidal *demodulation* [5] [10] (Figure 17). Synthesized carriers are multiplied by the local region followed by low-pass filtering to invoke the bandlimited constraint along the v-axis of the envelope terms in the GCT domain (30). Demodulation is done locally across time-frequency regions via a least-squared-error fitting method. The reconstructed spectrogram is combined with the phase of the original signal to estimate a waveform using overlap-add. This waveform estimate represents an "upper limit" of reconstruction due to inclusion of the phase of the original signal.

To obtain carrier parameters for voiced speech, we use the pitch mapping (31) in conjunction with prior pitch information. In contrast to the narrowband model, note that a *direct* mapping *forces all carriers to be located on the v-axis.* For unvoiced speech and in the *bootstrapping* method to be subsequently discussed, peak-picking is done using a multi-peak picker similar to that of [10]. The GCT magnitude is analyzed by a series of binary masks to extract peak locations based on a point's neighbors and amplitude thresholding.

Carrier assignments for demodulation are made for voiced speech using a *direct* method with mapped locations (for voiced speech). In the *bootstrapping* method, directly mapped carrier locations are reassigned to those obtained from peak-picking using a minimal distance criterion in an iterative algorithm (see Section III.C of [10]). Noise carriers are assigned based on peak-picking in both direct and bootstrapping approaches.

B. *Co-channel Speaker Separation*

As mentioned in the previous section, one motivation for analysis/synthesis with recovering the near-DC terms from their modulated versions is the separation (or removal) of interfering speakers. Specifically, we assume according to our model that near-DC terms of multiple speakers overlap, while carrier terms often do not, and that recovery of the (uncorrupted) DC region must be consistent with modulation of the carriers.

WGCT-Approach: In our WGCT-based approach, we apply least-squared-error demodulation using the *sum* of two modulation models to fit local time-frequency regions of the mixture spectrogram; as in [10] [5], this framework utilizes a sum-of-magnitudes approximation to the mixture spectrogram. Diagonal loading of the resulting least-squares matrix was performing using a threshold value obtained from a held-out development set [10]. Carrier parameters are obtained as in the single-speaker case using direct mapping and peak-picking. Permutations of mixture voicing conditions are used to assign carriers to distinct speakers for demodulation [10]. In the *voiced on voiced* case, the pitch mapping of (31) is used to obtain carrier positions that are used *directly* or as reference

values for *bootstrapping*/reassignment as in the single-speaker case using candidates from the peak-picker. In the *voiced on unvoiced* case, the *direct* pitch mapping is used to obtain the voiced speaker's carriers while the unvoiced speaker is assigned to carrier locations from peak-picking. In *bootstrapping*, the voiced speaker's carriers are first reassigned while the remaining candidate carrier locations from peak-picking are assigned to the unvoiced speaker. Finally, in the *unvoiced on unvoiced* case, carrier positions from peak-picking are used to fit the local region; the resulting estimate is halved and assigned to both speakers. A distinction of the WGCT approach from the NGCT is that we apply *bootstrapping* of the carrier positions as an alternative method to the *direct* approach instead of the exclusion/re-estimation method described in [10].

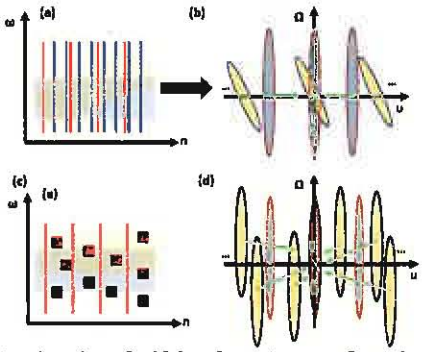

Figure 18 (a) Local region of wideband spectrogram for voiced speaker1 (red lines, shaded blue) and voiced speaker2 (purple lines, shaded yellow) mixture; (b) corresponding WGCT with removal of near-DC terms and demodulation to extract speaker1; (c) voiced speaker1 (red lines, blue shaded) and unvoiced speaker2 (black squares, yellow shaded) mixture; (d) WGCT of (c) indicating removal of near-DC terms and demodulation to recover speaker2. Note that demodulation in (b) and (d) are illustrated for the *direct* approach though this is done similarly in bootstrapping.

Fusion Methods: From Section VI, recall that the *number* of harmonic components in the GCT depends on short-time analysis window. For instance, male speakers with low pitch exhibit fewer terms in the NGCT than females while the opposite is true for WGCT; this effect was suggested in [10] as contributing to differences in performance in both analysis/synthesis and separation. It is conceivable that a *fusion* of separation estimates from both the NGCT and WGCT can lead to better overall estimates. We consider a simple fusion method using a weighted sum (here, $0 \leq \alpha \leq 1$)

$$\hat{x}_{fused}[n] = \alpha\hat{x}_{narrow}[n] + (1-\alpha)\hat{x}_{wide}[n]. \quad (43)$$

VIII. Evaluation

A. *Data Set*

In spectrogram analysis/synthesis and speaker separation, we use data from TIMIT [15] identical to that in [10]. For analysis/synthesis, 10 males and 10 females speaking 2 distinct utterances are used for a total of 40 examples. In separation, the development set consists of 5 male-male (MM), female-male (FM), and female-female (FF) mixtures while the test set consists of 24 male-male (MM), 24 female-female (FF), and 64 female-male (FM) mixtures, all mixed at 0 dB overall signal-to-signal ratio. The selected pairs cover a large range of overlapping voiced and unvoiced conditions, including crossing pitch tracks. Pitch tracks for individual utterances are obtained using the Wavesurfer package [16].

B. *Spectrogram Analysis/Synthesis*

Wideband spectrograms ($s_{full}[n,\omega]$) are computed as in Section III. GCT analysis is done using a 2-D 512-point DFT on local time-frequency regions of size 500 Hz by 37.5 ms extracted using a 2-D Hamming (overlap factor 4). A high-pass (low-pass) 1-D filter $h_{hp}[n]$ ($h_{lp}[n]$) is designed using the frequency sampling method $h_{hp}[n]$ ($h_{lp}[n]$) of order 80 with pass-band (stop-band) beginning at

$$0.5v_b = 60\frac{2\pi(0.625\times10^{-3}\times16000)}{16000} = 0.075\pi \quad (44)$$

corresponding to an extremal low-pitch case of 60 Hz from (31) (Section II.C) with stop-band (pass-band) roll-off to v_b. $h_{hp}[n]$ is applied to $s_{full}[n,\omega]$ to obtain $s_{full,hp}[n,\omega]$. $s_{full,hp}[n,\omega]$ is multiplied by a set of sinusoidal carriers followed by low-pass filtering by $h_{lp}[n]$ to obtain envelope estimates which are used to fit gain parameters in a least squares formulation. Note in demodulation, we used $s_{full,hp}[n,\omega]$ instead of $s_{full}[n,\omega]$; this was observed in preliminary experiments to reduce the influence of cross terms near WGCT origin after demodulation such as in the case of low-pitch values (e.g., for males).

As metrics, we compute root-mean-squared errors (RMSE)

$$RMSE = \sqrt{\frac{1}{N\omega_N}\sum_{n=1}^{N}\sum_{\omega=1}^{\omega_N}\left[s'_{full}[n,\omega] - \hat{s}'_{full}[n,\omega]\right]^2} \quad (45)$$

where ω_N denotes the total number of DFT frequency bins in the spectrogram and $s'_{full}[n,\omega]$ and $\hat{s}_{full}[n,\omega]$ are the original and reconstructions, respectively, normalized to have maximum value of unity. In addition, we compute the signal-to-noise ratio (SNR)

$$SNR = 10log\left(\frac{\sum_n x^2{}_{single}[n]}{\sum_n\left[x_{single}[n]-\hat{x}_{single}[n]\right]^2}\right) \quad (46)$$

where $\hat{x}_{single}[t]$ is waveform estimated obtained in combining $\hat{s}_{full}[n,\omega]$ with the phase of the original signal [10][12].

Figure 22 shows results of a single female utterance. For display purposes, we display the reference and reconstructed spectrograms after taking them to the power of 0.5; for reference, we show also an "error" spectrogram computed as the square difference between the bootstrap and true spectrograms after normalization. One limitation of the demodulation approach (in both bootstrapping and direct methods) is a "smoothing" effect on onset/offset structure, presumably due to bandlimiting of the envelope term in the proposed modulation mode (Figure 22, time 750). In addition, both methods fail to capture aperiodic content such as at time 500 as may be due to glottalization [12]. For voiced speech,

the "enforcement" of periodic carriers and their use as guides for reassignment in bootstrapping are evidently insufficient to fully address these effects. Overall, however, the reconstruction results demonstrate that speech content is generally well-represented by the modulation model with errors values ranging from 7e-3 to 4e-2 on a scale of unity as the maximum value. Quantitatively, bootstrapping appears to modestly outperform the direct method (Table 2) using the RMSE metric. Nonetheless, this is not reflected in the resulting waveforms in SNR, presumably due to phase effects in reconstruction. In informal listening, (non-author) subjects did not distinguish waveform reconstructions between bootstrapping, direct, and the original.

Table 2 Average RMSE and SNRs for analysis/synthesis of spectrograms and standard errors.

	Direct	Bootstrapping
RMSE (Males)	4.35e-2 [9.54e-3]	6.80e-3 [5.01e-3]
RMSE (Females)	3.32e-2 [6.09e-3]	7.92e-3 [5.44e-4]
SNR (dB) (Males)	24.61 [0.46]	22.15 [0.18]
SNR (dB) (Females)	21.89 [0.39]	23.04 [0.43]

C. *Co-channel Speaker Separation*

In speaker separation, the mixed signal $x_{mix}[n]$ is analyzed with short-time and GCT parameters dentical to those in analysis/synthesis. We compute RMSE errors in the spectrogram estimate as in analysis/synthesis for applications such as pre-processing for speech recognition. For human listening, reconstructed spectrograms are combined with the phase of the mixed signal to obtain a waveform estimate. We denote waveform estimates as $\hat{x}_i[n]$, and we compute the signal to interferer ratio [10]

$$SNR_i = 10log\left(\frac{\sum_n x^2{}_i[n]}{\sum_n [x_i[n]-\hat{x}_i[n]]^2}\right) \qquad (47)$$

where $x_i[n]$ is the original (unmixed) utterance. In fusion, the α parameter was swept on the development set from 0 through 1 with a step size of 0.01; we used the exclusion method from [10] and the bootstrap method the narrow and wideband estimates. The α value corresponding to the highest average SNR across all waveforms was used in testing. We obtained a "best" value of $\alpha = 0.56$ to be applied in testing.

Figures 23 and 24 show the results of wideband based speaker separation. Demodulation is capable of suppressing harmonic content from an interferer (e.g., time 750 (1200) in Figure 23 (24)), thereby leading to separation of speakers. A limitation in separation can be observed in Figure 24 (time ~1200) where the onset of the target is poorly replicated in the estimate. As in analysis/synthesis, this is likely due to bandlimiting of the envelope term in demodulation. Quantitatively, separation can result in RMSEs on the order of 3e-2 (on a scale of unity as the maximum value) and 4~6 dB global SNR gains across all permutations of mixtures (Table 3, Table 4). In general, bootstrapping appears to provide modest gains over the direct method. In our fusion results, we obtain global SNR gains up to ~1 dB over either narrow or wideband estimates alone. In Figure 19, the narrowband estimate provides better estimates overall; however, the wideband estimate provides complementary information in better suppressing content from an interferer at time ~2.6e4. In informal listening, (non-author) subjects reported suppression of interfering speakers in all conditions with faithful reconstructions of the target. Fused estimates were reported to exhibit less "abrupt" insertions of interfering speakers, consistent with the overall gains observed in SNR.

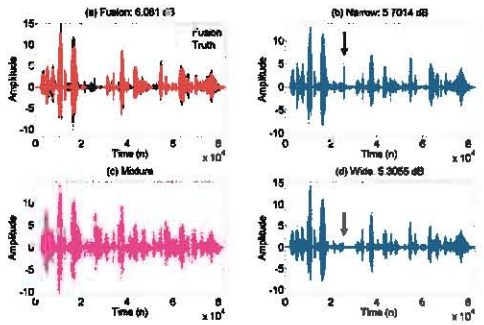

Figure 19. (a) Fusion estimate and truth target utterance "appetite"; (b) narrowband estimate of target; (c) mixture waveform of two females ("Neither his appetite, his exacerbations, nor his despair were akin to yours." + "Forty-seven states assign or provide vehicles for employees and state business.") (d) wideband estimate of target; note suppression in (b) of outstanding interferer in (d). fklh0.si1257.fmbg0.si1160.mix.wav

IX. Conclusions

This work has proposed a model of speech signal content as represented in 2-D analysis of wideband spectrograms. We have validated the utility of this model for representing speech content in both analysis/synthesis and co-channel speaker separation experiments. In conjunction with our previous work, the model motivates a novel *taxonomy* of speech signal behavior in the 2-D Grating Compression Transform (GCT) that exhibits important distinctions in interpretation, particularly in relation to "dual" behavior.

One implication of the proposed taxonomy is its potential for interpreting *other* time-frequency distributions. For instance, the auditory spectrogram of [1] is generally viewed as being "narrowband" / "wideband" in its low/high-frequency regions. The periodicity- and formant-dependent carrier derived in the current GCT framework may be applicable to high-frequency regions, thereby providing an explicit interpretation for modulation components observed in the auditory spectrogram in relation to speech parameters.

As suggested by our results in speaker separation, the GCT may have additional applications due to its representation of speech parameters. For instance, modifying carrier components in the WGCT may be used for pitch and/or formant bandwidth modification in voice transformation. As suggested in [10], the mapping of noise and speech content in distinct regions of the GCT space also motivates applicability to speech enhancement. Finally, the present speaker separation framework may be combined with existing multi-pitch tracking methods towards a full separation system.

Table 3 Average RMSEs for speaker separation and standard errors [] on test set.

	Direct	Bootstrap
MM	3.38e-2 [1.4e-3]	3.28e-2 [1.4e-3]
FF	3.22e-2 [8.96e-3]	3.19e-2 [8.81e-3]
FM – Male	2.77e-2 [8.61e-3]	2.82e-2 [9.29e-3]
FM – Female	3.52e-2 [1e-3]	3.64e-2 [1e-3]

Table 4 Average SNRs (dB) for speaker separation (dB), standard errors [] on test set.

	Direct	Bootstrap	Narrow	Fusion
MM	4.42 [0.12]	4.86 [0.15]	3.67 [0.16]	5.43 [0.16]
FF	5.63 [0.18]	6.02 [0.19]	6.30 [0.14]	6.72 [0.17]
FM – Male	5.46 [0.13]	5.92 [0.14]	4.83 [0.11]	6.51 [0.12]
FM – Female	5.66 [0.14]	5.54 [0.15]	5.71 [0.11]	6.49 [0.12]

APPENDIX I

Consider a *time-varying* decaying sinusoid represented by Green's function $g[n,m]$, where m is the time of excitation, and n is the time axis along which we observe the resulting response [12], i.e.,

$$g[n,m] = \xi e^{-\int_m^n \dot{\alpha}(z)dz} \cos\left(\int_m^n \dot{\phi}(z)dz\right) u[n-m]. \quad (48)$$

$\dot{\alpha}(z)$ and $\dot{\phi}(z)$ are integrable functions corresponding to the *instantaneous* decay rate and center frequency of the formant, respectively, and ξ is the initial amplitude of the response. The output $y[n]$ of $g[n,m]$ excited by $p[n]$ (4) is a superposition sum [12]

$$y[n] = \sum_{m=-\infty}^{\infty} g[n,m]p[m]. \quad (49)$$

Substituting (4) and (48) into (49), we obtain

$$y[n] = \sum_{k=0}^{N_k} \xi \begin{pmatrix} e^{-\int_{kP}^n \dot{\alpha}(z)dz} \cos\left(\int_{kP}^n \dot{\phi}(z)dz\right) \\ u[n-kP] \end{pmatrix}. \quad (50)$$

Let n_0 denote the time at which the window is shifted to extract a segment $y_{n_0}[n]$ of $y[n]$, i.e.,

$$y_{n_0}[n] = w[n-n_0] \sum_{k=0}^{N_k} \xi \begin{pmatrix} e^{-\int_{kP}^n \dot{\alpha}(z)dz} \cos\left(\int_{kP}^n \dot{\phi}(z)dz\right) \\ u[n-kP] \end{pmatrix}. \quad (51)$$

Consider n_0 in (51) such that the entirety of the window is located between impulses at N_kP and $(N_k+1)P$. Within $y_{n_0}[n]$, we assume that *the decay rate and frequency of the sinusoid are constant and a function of the time of analysis* n_0

$$y_{n_0}[n] \approx w[n-n_0] \sum_{k=0}^{N_k} \xi \begin{pmatrix} e^{-\left(\dot{\alpha}(n_0)n + \int_{kP}^{n_0} \dot{\alpha}(z)dz\right)} \\ \cos\left(\dot{\phi}(n_0)n + \int_{kP}^{n_0} \dot{\phi}(z)dz\right) u[n-kP] \end{pmatrix}. \quad (52)$$

This "frozen time" approximation is similar to that assumed in typical short-time analysis methods (e.g., linear prediction [17]) to invoke stationarity of speech parameters. The contribution of the k^{th} component in (52), although time-varying, appears to come from a decaying sine with constant decay and frequency. Nonetheless, its starting amplitude (of the decay) and phase (of the sinusoid) will differ as a function of the distance between n_0 and point of excitation kP.

We make a further approximation by assuming that each contribution to the summation across k in (52) is *aligned at the window onset such that it may be viewed as a scaled and shifted decaying sinusoid*, thereby ignoring effects of the phase terms $\int_{kP}^n \phi(z)dz$ and temporal overlap, i.e.,

$$y_{n_0}[n] \approx w[n-n_0] \sum_{k=0}^{N_k} e^{-\int_{kP}^{n_0} \dot{\alpha}(z)dz} h[n-n_0;n_0]. \quad (53)$$

$$h[n;n_0] = \xi e^{-\dot{\alpha}(n_0)n} \cos\left(\dot{\phi}(n_0)n\right) u[n] \quad (54)$$

The Fourier transform of (53) and its magnitude are

$$Y(n_0,\omega) \approx \sum_{k=0}^{N_k} e^{-\int_{kP}^{n_0} \dot{\alpha}(z)dz} [W(\omega) *_\omega H(\omega,n_0)] e^{-j\omega n_0} \quad (55)$$

$$H(\omega,n_0) = \frac{0.5\xi}{\dot{\alpha}(n_0)+e^{j(\omega-\dot{\phi}(n_0))}} + \frac{0.5\xi}{\dot{\alpha}(n_0)+e^{j\left(\omega+\dot{\phi}(n_0)\right)}} \quad (56)$$

$$|Y(n_0,\omega)|_{local} \approx w[n_0,\omega]E_d[n_0]\tilde{H}(n_0,\omega) \quad (57)$$

$$E_d[n_0] = \sum_{k=0}^{N_k} e^{-\int_{kP}^{n_0} \alpha(z)dz} \quad (58)$$

$$\tilde{H}(n_0,\omega) = |W(\omega) *_\omega H(\omega,n_0)|, \quad (59)$$

where $|\tilde{H}(n_0,\omega)|$ is a smoothed version of the formant and $E_d[n_0]$ is a time-dependent amplitude term. In (57), we have added a 2-D window term $w[n,\omega]$ to emphasize analysis in a local time-frequency region.

While $E_d[n_0]$ is not a periodic function in general, it can be made periodic in P under certain constraints such as $\dot{\alpha}(z) = \alpha_0$ or $\dot{\alpha}(z) = \cos\left(\frac{2\pi}{P}z\right)$ corresponding to constant or sinusoidally-varying decay rates. These conditions therefore allow for time-varying formants to be represented as a general time-dependent envelope term in conjunction with a periodic carrier. For instance, $\dot{\alpha}(z) = \alpha_0$ reflects a condition of constant decay but potentially changing formant frequency. For periodic $E_d[n_0]$, it can be shown that the 2-D Fourier transform of (57) (i.e., the WGCT) is

$$Y(v,\Omega) = W(v,\Omega) *_{v,\Omega} \left[\eta(v,\Omega) *_v \left(K\delta(v) + \sum_l^{N_l} 0.5\beta_l\delta\left(v \pm \frac{2\pi}{P}\right)\right)\right] \quad (60)$$

where $\eta(v,\Omega)$ is the 2-D Fourier transform of $\tilde{H}(n_0,\omega)$ and K, β_l, and N_l are parameters of a sinusoidal series.

Our discussion motivates a *modulation* view of the wideband spectrogram to include time-varying formant structure. Nonetheless, this view holds only approximately in time regions away from excitation impulse onsets due to the choice of the window position.

REFERENCES

[1] Chi, T., Ru, P. and Shamma, S., "Multiresolution Spectrotemporal Analysis of Complex Sounds." s.l. : Journal of Acoustical Society of America, 2005, Vol. 118.

[2] Jeon, W. and Juang, B., "Speech Analysis in a Model of the Central Auditory System." s.l. : IEEE Transactions on Audio, Speech, and Language Processing, 2007, Issue 6, Vol. 15.

[3] Greenberg, S. and Kingsbury, B., "The Modulation Spectrogram: In Pursuit of an Invariant Representation of Speech." Munich, Germany : Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997.

[4] Wang, T.T. and Quatieri, T.F., "High-pitch Formation Estimation by Exploiting Temporal Change of Pitch." s.l. : IEEE Transactions on Audio, Speech, and Language Processing, 2010, Issue 1, Vol. 18.

[5] Wang, T.T. and Quatieri, T.F., "Towards Co-channel Speaker Separation by 2-D Demodulation of Spectrograms." New Paltz, NY : roceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2009.

[6] Wang, T.T. and Quatieri, T.F., "Multi-Pitch Estimation by a Joint 2-D Representation of Pitch and Pitch Dynamics." Makuhari, Japan : roceedings of the 11th Annual Conference of the International Speech Communication Association, 2010.

[7] Quatieri, T.F., "2-D Processing of Speech with Application to Pitch Estimation." Denver, CO : Proceedings of the International Conference on Spoken Language Processing, 2002.

[8] Schimmel, S. and Atlas, L., "Coherent envelope detection for modulation filtering of speech." Philadelphia, PA : Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.

[9] Wang, T.T., *Exploiting Pitch Dynamics for Speech Spectral Estimation Using a Two-Dimensional Processing Framework.* Cambridge, MA : SM Thesis, MIT Department of Electrical Engineering and Computer Science, 2008.

[10] Wang, T.T. and Quatieri, T.F., "Two-dimensional Speech Signal Modeling." s.l. : in review, IEEE Transactions on Audio, Speech, and Language Processing.

[11] Nam, J, et al., "A Super-Resolution Spectrogram Using Coupled PLCA." Makuhari, Japan : roceedings of the 11th Annual Conference of the International Speech Communication Association, 2010.

[12] Quatieri, T.F., *Discrete-time Speech Signal Processing: Principles and Practice.* Upper Saddle River : Prentice Hall, Inc., 2001.

[13] Stevens, K.N., *Acoustic Phonetics.* Cambridge, MA : MIT Press, 1999.

[14] Malyska, N and Quatieri, T.F., "Spectral Representations of Non-modal Phonation." s.l. : IEEE Transactions on Audio, Speech, and Language Processing, 2008, Issue 1, Vol. 16.

[15] Fisher, W., Doddington, G. and Goudie-Marshall, K., "The DARPA Speech Recognition Research Database: Specifications and Status." s.l. : Proceedings of the DARPA Workshop on Speech Recognition, 1986.

[16] , http://www.speech.kth.se/wavesurfer/. [Online]

[17] Makhoul, J., "Linear prediction: a tutorial review." s.l. : Proceedings of the IEEE, 1975, Issue 4, Vol. 63.

[18] Oppenheim, A. and Schafer, R., *Digital Signal Processing.* Englewood Cliffs, NJ : Prentice Hall, Inc., 1975.

[19] Van Trees, H., *Detection, Estimation, and Modulation Theory, Part I.* New York, NY : Wiley, 1968.

[20] Wu, M., Wang, D.L. and Brown, G., "A Multi-pitch Tracking Algorithm for Noisy Speech." s.l. : IEEE Transactions on Audio, Speech, and Language Processing, 2003, Vol. 11.

[21] Lim, J., *Two-Dimensional Signal and Image Processing.* Englewood Cliffs, NJ : Prentice Hall, 1990.

[22] Wang, T.T. and Quatieri, T.F., "2-D Processing of Speech for Multi-Pitch Analysis." Brighton, UK : Proceedings of the 10th Annual Conference of the International Speech Communication Association, 2009.

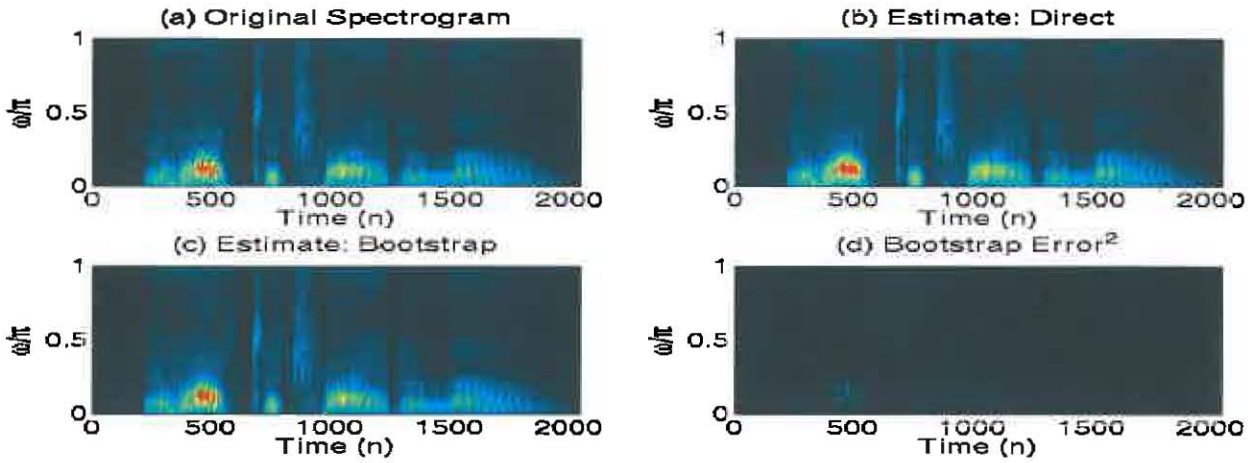

Figure 22. (a) Original spectrogram of female utterance "You'll have to try it alone."; (b) reconstruction using direct method; (c) reconstruction using bootstrapping method; (d) "Error" spectrogram computed as the square difference between (b) and (a). fceg0.si1878.0.specs.mat

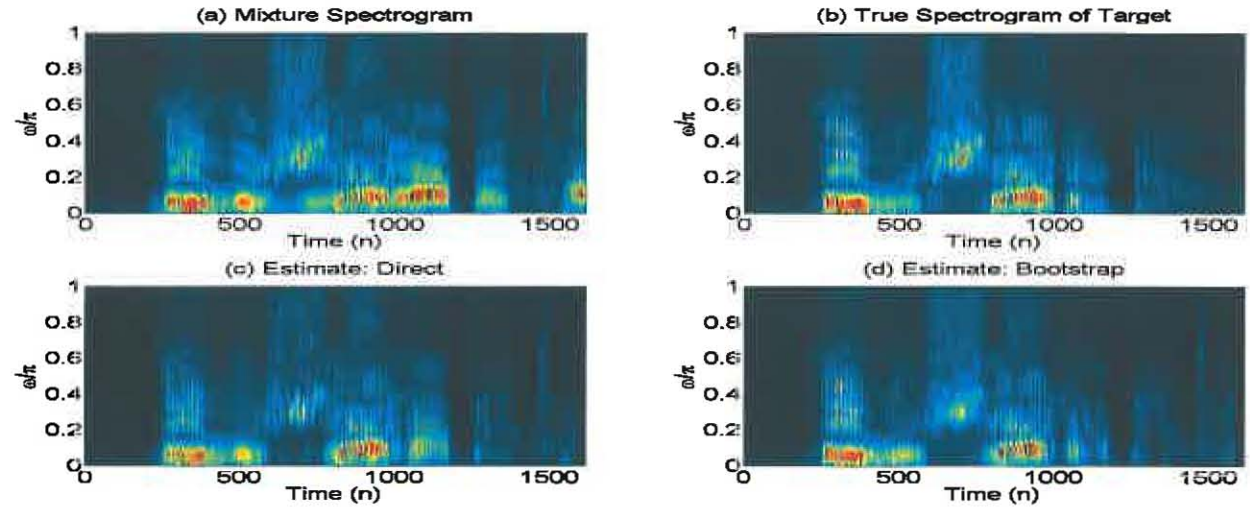

Figure 23. (a) Mixture spectrogram of a male ("They were shattered.") and female ("Neither his appetite"); (b) true male target; (c) male estimate using direct method; (d) male estimate using bootstrap method. For display purposes, spectrograms displayed are taken to the 0.5 power

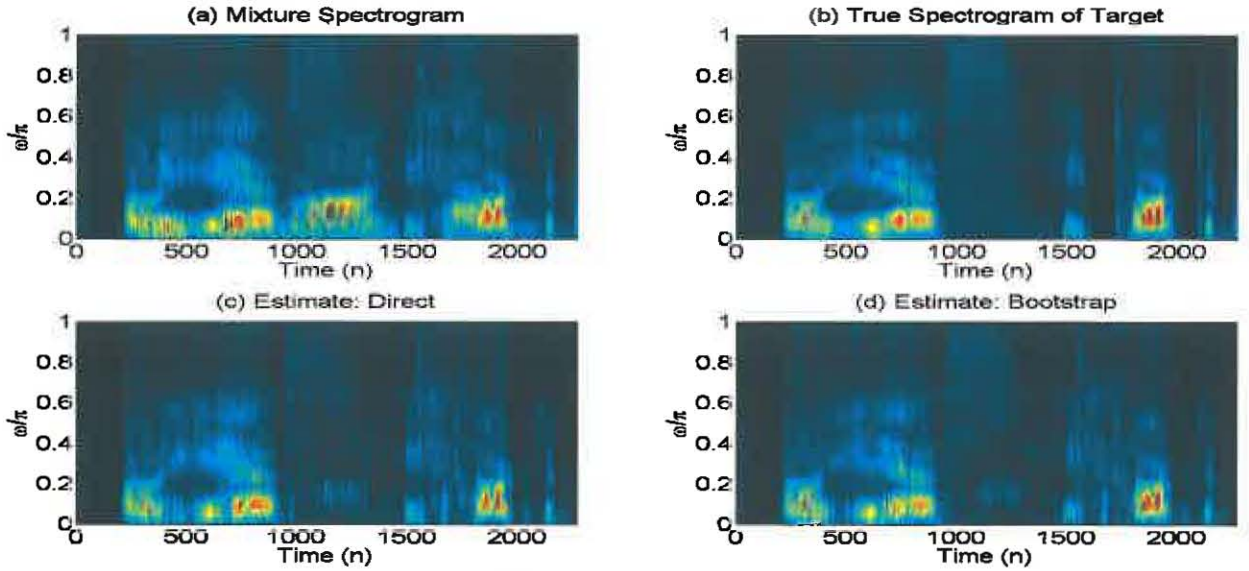

Figure 24. As in Figure 23 but for two female mixtures ("Oh yes, he talked", "Anything wrong captain?") with "talked" target utterance.